



HAL
open science

Development of a tool to assess the methodological quality of studies reporting on archaeologically excavated human skeletons: An international Delphi study

Amit Arvind Rajbhoj, Renée Speyer, Isabelle Crevecoeur, Giacomo Begnoni, Guy Willems, Maria Cadenas de Llano-Pérula

► To cite this version:

Amit Arvind Rajbhoj, Renée Speyer, Isabelle Crevecoeur, Giacomo Begnoni, Guy Willems, et al.. Development of a tool to assess the methodological quality of studies reporting on archaeologically excavated human skeletons: An international Delphi study. *Archaeometry*, In press, <10.1111/arc.12786>. <hal-03796822>

HAL Id: hal-03796822

<https://hal.science/hal-03796822v1>

Submitted on 5 Oct 2022



HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Development of a tool to assess the methodological quality of studies reporting on archaeologically excavated human skeletons: An international Delphi study

Amit Arvind Rajbhoj¹  | Renée Speyer^{2,3,4}  | Isabelle Crevecoeur⁵ | Giacomo Begnoni¹ | Guy Willems¹ | Maria Cadenas de Llano-Pérula¹

¹Department of Oral Health Sciences—Orthodontics, KU Leuven and Dentistry, University Hospitals Leuven, Leuven, Belgium

²Department Special Needs Education, University of Oslo, Oslo, Norway

³School of Occupational Therapy, Social Work and Speech Pathology, Faculty of Health Sciences, Curtin University, Perth, Western Australia, Australia

⁴Department of Otorhinolaryngology and Head and Neck Surgery, Leiden University Medical Centre, Leiden, The Netherlands

⁵University of Bordeaux, CNRS, MCC, PACEA, UMR 5199, Pessac, France

Correspondence

Amit Arvind Rajbhoj, Department of Oral Health Sciences—Orthodontics, KU Leuven and Dentistry, University Hospitals Leuven, Kapucijnenvoer 7, 3000 Leuven, Belgium.
Email: amitarvind.rajbhoj@kuleuven.be

Abstract

Methodological bias can directly affect the interpretation of research data. Studies reporting on excavated skeletons represent a valuable source of information in medicine, dentistry, archaeology and anthropology, and forensic sciences. However, these studies represent a specific setting with their own methodology, for which no quality assessment tool is available. The aim was to develop a critical appraisal tool to assess the methodological quality of studies reporting on archaeologically excavated human skeletons. An international Delphi study was therefore conducted to support item generation and ensure content validity for a new tool. Experts from the following domains were consulted: dentistry, forensic sciences, archaeology and anthropology, general medicine, epidemiology, and statistics. Participants judged the relevance and comprehensiveness of items retrieved from the literature. Consensus was predefined as 75% agreement between experts, and achieved within two Delphi rounds. As a result, 44 and 32 participants completed the first and second Delphi rounds, respectively, achieving consensus on 17 items.

This research provides the first evidence-based tool for the methodological assessment of studies reporting on archaeologically excavated skeletons. Clinicians and researchers can use this tool for critical appraisal of studies or when performing systematic reviews. Future research will focus on psychometric testing of the newly developed tool.

KEYWORDS

anthropology, Delphi study, evidence, human skeletal remains, risk of bias, systematic review, tool

INTRODUCTION

Human remains can provide ample information to researchers within the domains of archaeology, anthropology, forensic sciences, medicine, and dentistry. This evidence represents a source of valuable knowledge. However, flaws in study design, methodology, or analysis may result in bias, compromising evidence interpretation (Beck & Jones, 1989; Higgins et al., 2011; Jackes, 2011).

Although guidelines for standardizing the reporting of observational studies are available (e.g., STROBE, STrengthening the Reporting of OBServational studies in Epidemiology; von Elm et al., 2014), in some fields, such as paediatric surgery (Rangel et al., 2003) or genetics (Al-Jader et al., 2002), domain-specific tools have been developed for evaluating those observational studies that cannot be assessed using more generic quality assessment tools (Coles et al., 2021). No critical appraisal tool (CAT) is available to address the methodological quality of studies reporting on archaeologically excavated skeletons. Existing tools designed for assessment of observational or interventional trials in biomedical literature, such as those that can be found in www.equator-network.org, are not applicable to these specific kinds of studies, due to their unique potential sources of bias related to skeleton representativeness, preservation, aging, sex determination, and ancestry (Jackes, 2011; Walker et al., 1988). Also, specific guidelines in the fields of archaeology and anthropology, such as those proposed by Nikita and Karligkioti (2019), Fladmark (1978), Mays et al. (2004), Buikstra and Ubelaker (1994), IFA paper no. 7 published by BBAO updated in 2017 (Brickley & McKinley, 2017), or Field guide 14 of the British Archaeological Resource (OSSAFreelance, 2005), are to be considered during skeletal excavation procedures.

The development of a measurement instrument involves a systematic procedure that usually starts with the process of item generation, which provides theoretical support for the initial list of items to be included in the tool. Next, the preliminary measure may be trialled in the target population, after which its psychometric properties will be determined using robust statistical analyses. Based on these analyses, a measure may be revised, trialled, and psychometrically tested again. The COSMIN group (COnsensus-based Standards for selection of health Measurement INstruments) (Mokkink, Terwee, Patrick, et al., 2010) established international consensus-based taxonomy, terminology, and definitions of measurement properties. The framework comprises nine measurement properties within three domains: reliability, validity, and responsiveness. Content validity is the most important psychometric property when selecting an instrument and ensures adequate reflection of the construct to be measured through theoretical analysis (Mokkink, Terwee, Knol, et al., 2010). Good content validity can be warranted if a measure has been developed based on current literature and with reference to expert groups and target population focus groups (if appropriate).

Formal consensus on items retrieved from the literature can be achieved using the Delphi method (Mahajan et al., 1976; Morgado et al., 2018; Whiting et al., 2003). This is a formal expert consensus method that helps in assessing the three aspects of content validity (relevance, comprehensiveness, and comprehensibility of the items and instructions of the tool; Yoon et al., 2020). The Delphi technique was first used in the 1950s for defence research by the US Air Force. Their objective was to obtain reliable consensus on the advice given by a group of experts, which required an intensive series of questionnaires together with controlled opinion feedback (Mahajan et al., 1976). This technique later caught the attention of researchers outside the field of military defence and became a cornerstone in technology foresights and innovation, public policy making and development of measurement instruments such as risk of bias tools (Diamond et al., 2014; Terwee et al., 2018).

In the preliminary conceptualization phase of construction of a quality assessment tool the following characteristics should be considered (Morgado et al., 2018; Whiting et al., 2003, 2017):

1. It should be short, simple to understand, and intuitive to apply.
2. It should determine the methodological quality of the studies, making it possible to easily distinguish between studies with high and low risk of bias.

In an effort to perform a systematic review on dental occlusion involving excavated skeletons, we were unable to find any tool to assess the methodological quality of such studies. The lack of such a tool can make both systematic reviews and guidelines for complete analysis of archaeological skeletons difficult. Hence the aim of this study was to systematically develop a tool to assess the methodological quality of studies that report on archaeologically excavated human skeletons by using the Delphi formal consensus method. The present study focuses on the first two steps of tool development, namely item generation and content validity.

MATERIALS AND METHODS

This study follows the framework for construction of quality assessment tools proposed by Morgado et al. (2018) and Whiting et al. (2017).

Item generation

To develop the initial list of questions, we searched the literature using the terms ‘excavated skeletons’, ‘burials’, ‘archaeology’, and ‘physical anthropology’ in the electronic database PubMed, from inception until the 23 October 2020. To identify evidence for potential sources of bias in this particular sort of study, we added the terms ‘bias’ and ‘quality assessment’. Based on these terms, we obtained and reviewed a list of studies that included excavated human skeletons. Various guidelines concerning field excavations were also consulted (Brickley & McKinley, 2017; Fladmark, 1978; Mays et al., 2004; Nikita & Karligkioti, 2019; OSSAFreelance, 2005) and systematic reviews in the field of archaeology and anthropology were hand-searched in order to find possible assessment tools. We did not find any previous research concerning methodological quality assessment tools of studies reporting on archaeologically excavated human skeletons.

A possible list of 17 initial items was formulated by one researcher in the team (AA). The preliminary list of items was based on (1) the outcome of an ongoing systematic review on archaeologically excavated skeletons, (2) the literature from the PubMed database searches, and (3) the field guidelines (Brickley & McKinley, 2017; Fladmark, 1978; Mays et al., 2004;

Nikita & Karligkioti, 2019; OSSAFreelance, 2005). This initial list was discussed with another reviewer—an expert in the same field (MC)—after which a first proposal was electronically made available to two additional researchers (IC, RS), belonging to different fields of expertise (anthropology and archaeology, epidemiology and logopedics). Feedback regarding item formulation, relevance, and scoring criteria was asked. Disagreements were resolved by reaching a consensus. The initial list of items is shown in Supporting Information Annex A. The process of tool development for the assessment of the methodological quality of studies reporting on archaeologically excavated skeletons is shown in the flowchart of Figure 1.

Item scoring was inspired by a study from Slim et al. (2003), developed for the Methodological Index of Non-Randomized Studies (MINORS). The MINORS tool provides a numeric score, which can be formally incorporated in the systematic review.

Assessment of content validity: The Delphi study

As mentioned in the Introduction, content validity provides information on the relevance of the construct of interest as well as comprehensiveness and comprehensibility of the items in the tool. We used the Delphi formal consensus method to warrant sufficient content validity of our new measure on studies reporting on archaeologically excavated human skeletons.

The Delphi technique is a method for structuring a group communication process that aims for consensus on a defined topic through a series of questionnaires or survey rounds along with controlled feedback (Denman et al., 2019; Mahajan et al., 1976). This technique was used in the present study to obtain consensus among a group of experts on the items to be included in the tool, the phrasing of the items, and their scoring method. As the expert anonymously completes each survey round, the responses of each expert are secured without any influence from others (Hasson et al., 2000). The same experts continue the subsequent rounds until a group consensus of at least 75% is obtained or until it is sure that consensus cannot be achieved on one or more of the items included in the initial list of questions (Diamond et al., 2014). The results of the previous round along with the suggested revisions are used to further stimulate opinion building and consensus between the experts in subsequent rounds.

Participant characteristics

Experts from the domains of dentistry (156), forensic sciences (116), archaeology and anthropology (106), general medicine (100), and epidemiology and statistics (108) were consulted, since the tool could be used in a broad spectrum of fields. We attempted to achieve sufficient representation of experts from each domain, so that the tool could be applicable to multiple fields. ‘Experts’ were defined as individuals who are knowledgeable in the subject under consideration. They were included based on a minimum qualification of post-graduation, research activity as shown by recent scientific publications, or involvement in academic teaching at university level. International experts were included without any geographic limitations. The experts were contacted via email by searching their contact details from published studies in the electronic database PubMed, university websites, national and international conference contact lists, or through the professional networks of the authors. The email included a brief introduction, the procedure and link to the survey, as well as the consent to participate, where it was specified that their participation would remain anonymous (Annex B). Experts who failed to complete the questionnaire in the first survey rounds were excluded from the second one. The participant demographics are shown in Table 1.

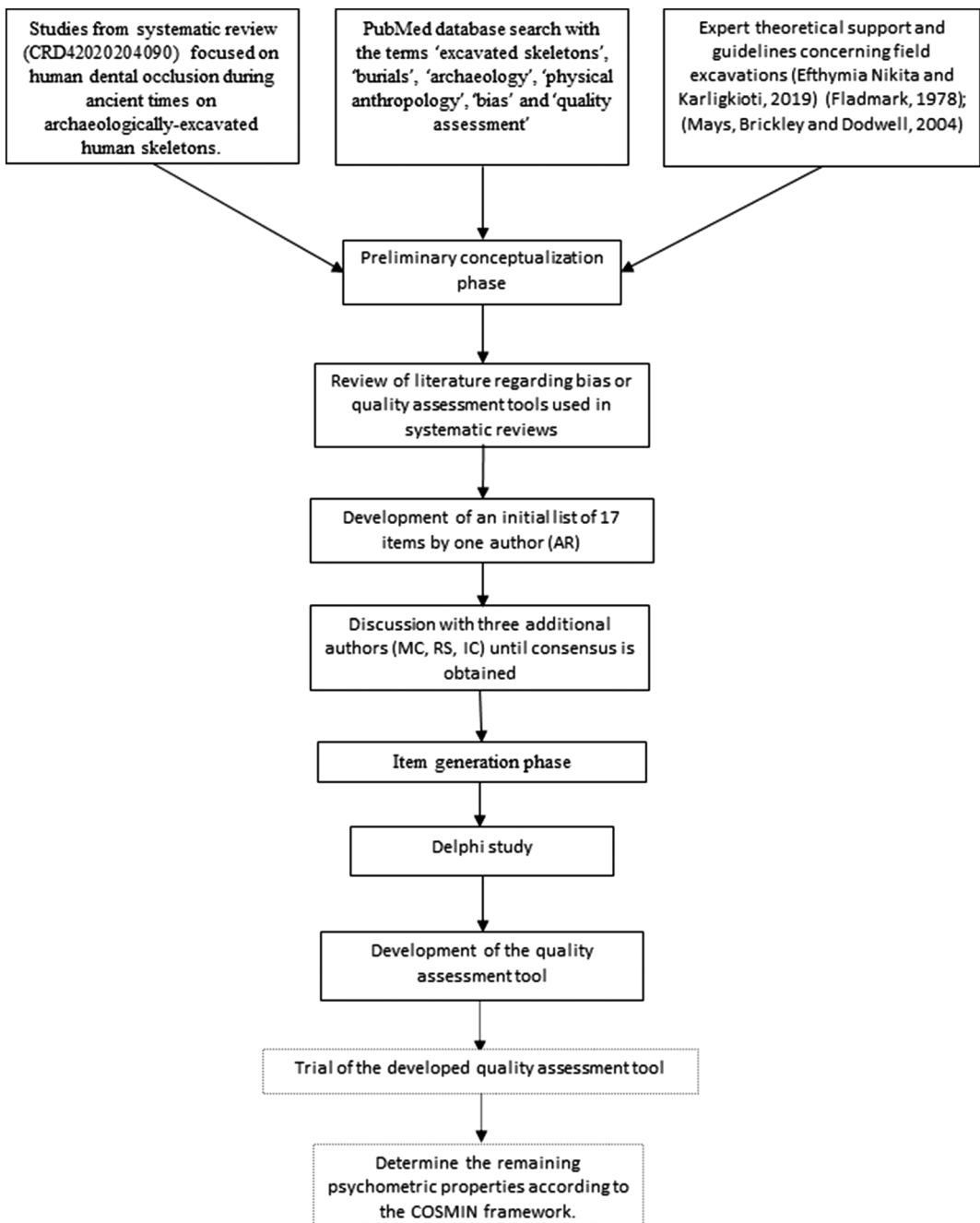


FIGURE 1 Flowchart of the development of a tool for the assessment of the methodological quality of studies reporting on archaeologically excavated skeletons. The steps in the dotted boxes are to be completed in order to validate and generate the refined measurements, which will be done in another study

Delphi round one

The pilot version of the tool including 17 items was presented to the experts through an online survey using Qualtrics software (XM Solutions, USA). A statement with a brief explanation of

TABLE 1 Participant demographics

Category	Subcategory	Round one <i>n</i> (%)	Round two <i>n</i> (%)
Country	Australia	3 (6.8%)	3 (9.3%)
	Belgium	8 (18.2%)	7 (21.9%)
	China	1 (2.3%)	1 (3.1%)
	Finland	4 (9.1%)	4 (12.5%)
	France	1 (2.3%)	1 (3.1%)
	Ireland	1 (2.3%)	0 (0%)
	India	6 (13.6%)	4 (12.5%)
	Italy	1 (2.3%)	1 (3.1%)
	Norway	2 (4.5%)	1 (3.1%)
	New Zealand	1 (2.3%)	1 (3.1%)
	Romania	1 (2.3%)	0 (0%)
	Sweden	1 (2.3%)	1 (3.1%)
	Switzerland	1 (2.3%)	1 (3.1%)
	Spain	2 (4.5%)	1 (3.1%)
	UK	3 (6.8%)	3 (9.4%)
USA	8 (18.2%)	3 (9.4%)	
	Total	44	32
Field of expertise	Anthropology & Archaeology	6 (13.6%)	5 (15.6%)
	Dentistry	16 (36.4%)	14 (43.8%)
	Forensic Sciences	14 (31.8%)	9 (28.1%)
	Epidemiology/Statistics	3 (6.8%)	2 (6.3%)
	General Medicine	5 (11.4%)	2 (6.3%)
		Total	44

n, number of experts participating.

the variable of interest for each item was provided (regarding reporting, internal and external validity). Experts were asked to rate (1) their agreement on the given statements, (2) the importance of the variable in the context of studies reporting on archaeologically excavated skeletons, and (3) the scoring method of each variable. The level of agreement for all three steps was scored on a 5-point Likert scale (5: strongly agree; 4: agree, 3: neither agree nor disagree; 2: disagree; 1: strongly disagree). The experts could comment or suggest any possible rephrasing or modification for each item and specify the rationale concerning their rating. They could suggest any additional items if necessary. The time needed to complete each survey round was estimated to be 30 min after conducting a pilot survey by two experts from different domains. A period of 1 month was provided for the first round; a reminder email was sent 2 weeks after the initial invitation to participate (Hsu & Sandford, 2007). The complete survey can be found in Supporting Information Annex A and Annex C.

Delphi round two

Descriptive statistics were determined for round one: percentages, medians, and interquartile ranges. The comments from the experts were implemented after discussion and agreement among authors, and changes made were highlighted in the next round. If at least 75% of the

experts responded 'strongly agree' or 'agree' to a specific item, these were included in the final tool (Denman et al., 2019; Diamond et al., 2014). In round two, we asked the experts to rate the revised items for which no consensus was achieved and also provided them with the opportunity to decide whether to change their responses or remain with the initial decision. The experts were also asked again to comment or suggest any modifications for each item and to specify their rationale concerning their rating. The descriptive statistics of round one were made available for the experts in round two, so that they could situate their answers within those of the group.

Statistical analysis

Responses were analysed using the Statistical Package for Social Sciences (SPSS version 20, IBM Corp.). As suggested in the literature, a group response of 'strongly agree' or 'agree' along with a percentage agreement of 75% or more was considered as strong consensus (von der Gracht, 2012; Denman et al., 2019; Diamond et al., 2014).

A median of ≥ 4 and an interquartile range (IQR) of ≤ 1 report a strong consensus. Medians were used since data were ranked and not scaled, and medians have been reported to be particularly useful in forecasting whether data contain outliers (Goodwin et al., 2002; von der Gracht, 2012). Means were not considered since outliers can affect them unrealistically. In the Delphi method, measure of central tendency should be analysed in connection with dispersion (as this indicates the spread of the scores). Range is the simplest measure that can be used for this. However, as range changes with extreme values, interquartile ranges were used to compensate for this effect, since they are an objective way of determining consensus (Murphy et al., 1998).

RESULTS

An invitation link was sent to 586 participants via email. From these, 53 experts agreed to participate in the survey (acceptance rate 9.04%). Forty-four of them completed the first round, yielding a response rate of 83.01%. Thirty-two of the initial 44 completed round two (response rate of 72.72%). A flowchart regarding the recruitment process is presented in Figure 2.

Experts from 16 countries participated in the first round (range: 1–8 experts per country). Most experts were from Europe (25), followed by North America (8), Asia (7), and Oceania (4); see Table 1.

Agreement on statements

For the provided statements in round one, consensus was obtained on all 17 items in the tool, with a percentage agreement of more than 84% (median ≥ 4 , IQR ≤ 1). Seven of the 17 variables showed a median of 5, while one ('Method[s] of measurement') reported an IQR of 0 and a percentage agreement of 100% (Table 2).

Importance of variables

In round one, the experts reached consensus on 16 of the 17 items regarding their importance to be included in the tool. For 11 of these 16 variables, a median score of 5 on the Likert scale was reported, and one of the 16 variables ('Method of measurement[s]') reported an

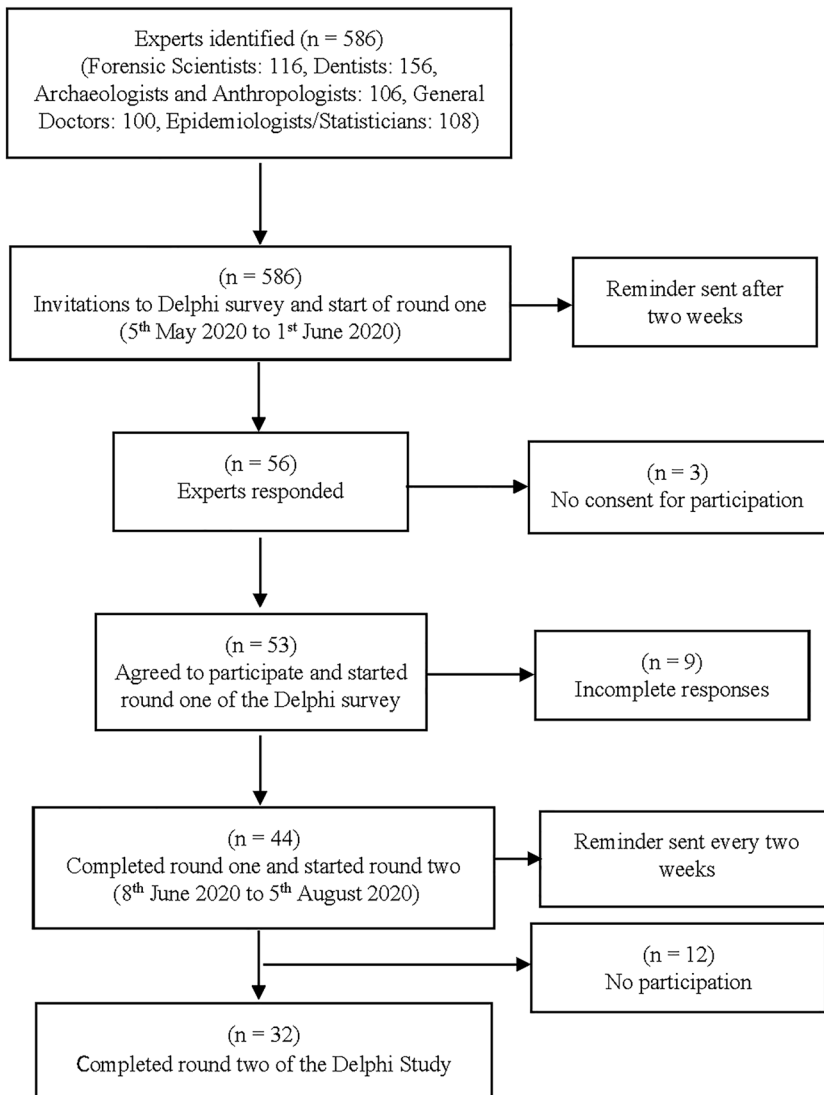


FIGURE 2 Flowchart of the recruitment process of the Delphi study

IQR of 0. For 16 of the 17 items, the achieved percentage agreement was strong, ranging from 77.2% ('Ancestry distribution') to 100% ('Method for Age-at-death estimation' and 'Method[s] of measurement'). Consensus was not achieved on one item ('Blind Testing'; median: 4; IQR: 2; agreement of 65.9%). After revision of this item, strong consensus was achieved during round two of the Delphi survey, with an agreement of 84.4% (median: 4; IQR: 1) (Table 2). All the experts who disagreed on the importance of 'Blind Testing' in round one (scores 1 and 2: strongly disagree and disagree) also participated in round two. 30.7% of the experts who found the variable 'Blind Testing' of limited importance (score 3) dropped out in round two (3 out of 14 experts).

TABLE 2 Assessment of the items in the tool in rounds one and two of the Delphi study

Item	Agreement on statement			Importance of variable for tool			Agreement on scoring of variable		
	Round one (n = 44)			Round one (n = 44)			Round one (n = 44)		
	M	IQR	%A	M	IQR	%A	M	IQR	%A
1	4	1	88.70%	5	1	93.10%	4	1	86.30%
2	4	1	88.70%	4	1	90.90%	4	1	88.60%
3	4	1	88.60%	5	1	77.20%	4	1	77.30%
4	5	1	97.70%	5	1	93.20%	4	1	88.70%
5	4.5	1	88.60%	4	1	90.90%	4	1	77.30%
6	5	1	100%	5	1	95.50%	4	1	90.90%
7				5	1	100%	4	1	90.90%
8	5	1	97.70%	5	1	95.50%	4	1	88.60%
9				5	1	95.50%	4	1	88.70%
10	5	1	93.10%	5	1	93.10%	4	1	88.60%
11				5	1	90.90%	4	1	86.30%
12	5	0	100%	5	0	100%	4	1	91%
13	5	1	97.70%	5	1	93.20%	4.5	1	81.80%
14	5	1	95.50%	4.5	1	93.20%	4	1	86.40%
15	4	1	84.10%	4*	2*	65.9%*	4*	2*	70.5%*
16	4	1	88.60%	4	1	81.80%	4	1	79.60%
17	4	1	93.20%	4	1	77.30%	4	0	81.80%

Abbreviations: IQR, interquartile range, i.e. middle 50% of the data (difference between 75th and 25th percentile); M, median, NA, not applicable, %A, percentage agreement, 1: clearly stated aim; 2, primary outcome; 3, racial/ethnic distribution; 4, condition of the sample; 5, sample preservation; 6: age-at-death; 7, method for age-at-death estimation; 8, sex distribution; 9, method for sex determination; 10, time period of the sample; 11, dating method; 12: method(s) of measurement; 13, statistical tests; 14, outcome reporting; 15, blind testing; 16, baseline equivalence of groups; 17: presence of other bias; *consensus not obtained.

Agreement on scoring of variables

The overall score for comparative articles and non-comparative articles is 34 and 32, respectively. Strong consensus on the scoring of the variable was achieved for 16 of the 17 items in round one (median: 4). IQR was '0' for the variable 'Presence of other bias' and 15 of the 17 items obtained an IQR of '1', except for the variable 'Blind Testing' (median: 4; IQR: 2). Consensus ranged from 77.3% ('Ancestry/Ethnicity distribution' and 'Sample preservation') to 91% ('Method[s] of measurement'). The percentage agreement for scoring was 70.5% for the variable 'Blind Testing'. Because the predefined consensus of 75% was not reached for one variable, we conducted a second round in which a consensus of 84.4% for Blind Testing was achieved after adaptation (Table 2). All the experts who disagreed on the scoring of the variable 'Blind Testing' in round one (scores 1 and 2: strongly disagree and disagree) also participated in round two. 36.3% of the experts who neither agreed nor disagreed on the scoring of the same variable (score 3) in round one dropped out in round 2 (4 out of 11).

Expert feedback

Phrasing of the items was the most commented issue by the experts. For example, experts suggested the use of the term 'findings', referring to study results and their interpretation, instead of results, which was seen more as a paper section. The term 'ancestry' was preferred instead of 'racial/ethnic', since in the study of human anatomy estimation of ancestry is often performed. The specification 'analysis not performed separately' was also added to the scoring of this particular item after round one, because the way heterogeneity is handled and reported determines the quality of the study. For the item 'age-at-death' articles were scored 1 point if the 'information was given only for part of the sample', as in some studies age estimation cannot be established for all samples. 'Date of the sample' was rephrased as 'time period' of the samples (Supporting Information Annex A and Annex C).

DISCUSSION

The present study provides the first evidence-based tool for assessment of the methodological quality of studies reporting on archaeologically excavated human skeletons. This tool can be used in systematic reviews or for critical appraisal of related studies by researchers, clinicians, reviewers, or readers. During the process of instrument development, the original focus was on studies within the field of dentistry, but the final tool has a broader scope. It can be applied to studies reporting on archaeologically excavated skeletons in forensic sciences, medicine, archaeology, and anthropology, among others.

The development of a valid and reliable quality assessment tool needs to meet several standards. The present tool was based on the Delphi method according to the COSMIN framework (Mokkink et al., 2020; Streiner, 2003; Wright, 1992). In the literature, it has been argued that this method improves content validity as it allows experts from different domains and with different viewpoints to structurally form a consensus on each item included in the tool (Whiting et al., 2003). The quality of the development process is enhanced by facilitating anonymity, iteration, and controlled feedback. In a Delphi study, the number of rounds and the results of the study reflect the consensus opinion. The optimal approach is to predefine the criteria of what constitutes consensus, and the most common definition of consensus is based on percentage agreement of 75% (Diamond et al., 2014).

Experts from five different domains were invited to discuss and agree upon the items as part of a critical appraisal tool to support methodological quality assessment of studies reporting on

the use of archaeologically excavated skeletons in these research areas (Mokkink, Terwee, Knol, et al., 2010). The presence of a heterogeneous expert panel might influence the responses across and between the specialties. However, it also ensures its use across fields and avoids field specificity to bias results. In this study, the pilot item list was accepted in round one, showing strong consensus with only one item lacking sufficient consensus, and the final list achieved consensus for all items in the tool. As a matter of fact, heterogeneity in expert groups has been reported to have a better performance than homogeneity (Bantel, 1993; Hong, 2010), which is especially relevant to our particular area due to its inherent multidisciplinary nature.

The initial acceptance rate of the survey was low, despite the efforts put into maximizing the response rate. The Dillman approach for internet-based surveys was used, which involves personalized repeated contact with the individual expert. This was accomplished by sending email reminders every 2 weeks in each Delphi round (Dillman et al., 2014). However, since no information regarding acceptance rate is available for similar studies, we could not compare our data with others. Although not uncommon, the further dropout of the experts after round one, even after the bimonthly reminders, could be considered as a shortcoming. This dropout could be due to the large number of questions included in the survey and the time required to complete it. All items that achieved consensus were adapted and included again in round two, which may have discouraged some experts. However, all experts who disagreed in round one participated in round two, which strengthens the final result. The minimum number of experts required has actually not been defined in Delphi literature, and the number of experts included in the present study is in line with that of previous reports (Al-Marzouki et al., 2005; Armstrong et al., 2005). As stated by Akins et al. (2005), responses from a small number of experts with a well-defined knowledge about the topic are stable. Also, it is important to remark that this was a web-based approach, which tends to produce a lower response rate, as previously reported by Boulkedid et al. (2011). The high response rate obtained from dentists might be related to a large number of these experts initially approached and the fact that the invitation email was sent by the first author, who is a dentist.

The newly developed critical appraisal tool not only involves reporting of the items but also emphasizes the conduct of the study. Items regarding preservation of excavated materials, estimation of age-at-death, determination of sex, methods of measurement, and description of the methods used, refer more to the study's setting and differ from those normally involved in biomedical research. These items are also part of recommended guidelines for field excavations (Brickley & McKinley, 2017; Fladmark, 1978; Mays et al., 2004; Nikita & Karligkioti, 2019; OSSAFreelance, 2005).

Bones and teeth are durable tissues often found in human archaeological excavations. Archaeological analysis of these tissues has contributed to answering questions concerned with evolution and proliferation of a number of diseases, ranging from tuberculosis to changes in the dentition and dietary habits. Hence the use of this tool for quality assessment of studies involving excavated skeletons will not only be limited to archaeologists or anthropologists but can also be applied to forensic sciences, medicine and dentistry (Donoghue & Rücklin, 2016; Evensen & Øgaard, 2007; Holloway et al., 2011; Kaidonis, 2008; Lavelle, 1972; Quam et al., 2009; Scott, 2018; Wood, 1996).

The scoring of the items was also evaluated. Cut-off points to categorize the studies as 'high risk' or 'low risk' for bias were avoided. This would ignore the importance of each item, since individual bias might vary according to the study design and the article's specific context. It is preferred to see how each trial was scored on each item. The overall score for each article can be interpreted as a whole by graphical plots to identify the risk of bias of the particular articles (Enthoven et al., 2016). In circumstances where some of the items of the tool are not relevant, these items can either be rated with the highest score (to prevent them from being wrongly interpreted as a high risk of bias), or simply be considered as not applicable and excluded from the

total score. This strategy has already been used in the literature by other risk-of-bias tools (Rajbhoj et al., 2021; Slim et al., 2003; Sterne et al., 2016).

Based on the current results of the Delphi study, the COSMIN framework will be used to evaluate the remaining psychometric properties of the instrument within the domains of reliability, validity, and responsiveness.

CONCLUSION

The present article proposes the first evidence-based tool for assessment of the methodological quality of articles reporting on archaeologically excavated human skeletons. This tool can be used in the fields of archaeology, anthropology, forensic sciences, dentistry, or medicine by researchers performing systematic reviews or by clinicians, editors, reviewers, or simply readers, since it enables standardized assessment of risk of bias and facilitates comparisons between studies. For conception and development of the tool, a formal expert consensus method (Delphi) was followed. The remaining psychometric properties according to the COSMIN framework will be evaluated separately.

ACKNOWLEDGEMENT

We wish to thank the 44 anonymous experts who gave their valuable time to this study.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/arcm.12786>.

DATA AVAILABILITY STATEMENT

The data from this project were collected as a part of the corresponding authors dissertation to allow for publishing in the dissertation chapters. However, in support of open science, data can be made available on request by an email to the corresponding author: amitarvind.rajbhoj@kuleuven.be

ORCID

Amit Arvind Rajbhoj  <https://orcid.org/0000-0002-4932-2428>

Renée Speyer  <https://orcid.org/0000-0003-2828-8897>

REFERENCES

- Akins, R. B., Tolson, H., & Cole, B. R. (2005). Stability of response characteristics of a Delphi panel: application of bootstrap data expansion. *BMC Medical Research Methodology*, 5(1), 37. <https://doi.org/10.1186/1471-2288-5-37>
- Al-Jader, L., Newcombe, R. G., Hayes, S., Murray, A., Layzell, J., & Harper, P. S. (2002). Developing a quality scoring system for epidemiological surveys of genetic disorders. *Clinical Genetics*, 62(3), 230–234. <https://doi.org/10.1034/j.1399-0004.2002.620308.x>
- Al-Marzouki, S., Roberts, I., Marshall, T., & Evans, S. (2005). The effect of scientific misconduct on the results of clinical trials: A Delphi survey. *Contemporary Clinical Trials*, 26(3), 331–337. <https://doi.org/10.1016/j.cct.2005.01.011>
- Armstrong, D., Marshall, J. K., Chiba, N., Enns, R., Fallone, C. A., Fass, R., Hollingworth, R., Hunt, R. H., Kahrilas, P. J., Mayrand, S., Moayyedi, P., Paterson, W. G., Sadowski, D., van Zanten, S., & Canadian Association of Gastroenterology GERD Consensus Group. (2005). Canadian Consensus Conference on the Management of Gastroesophageal Reflux Disease in Adults – Update 2004. *Canadian Journal of Gastroenterology*, 19(1), 15–35. <https://doi.org/10.1155/2005/836030>
- Bantel, K. A. (1993). Comprehensiveness of Strategic Planning: The Importance of Heterogeneity of a Top Team. *Psychological Reports*, 73(1), 35–49. <https://doi.org/10.2466/pr0.1993.73.1.35>
- Beck, C., & Jones, G. T. (1989). Bias and Archaeological Classification. *American Antiquity*, 54(2), 244–262. <https://doi.org/10.2307/281706>

- Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and Reporting the Delphi Method for Selecting Healthcare Quality Indicators: A Systematic Review. *PLoS ONE*. Edited by J. M. Wright, 6(6), e20476. <https://doi.org/10.1371/journal.pone.0020476>
- Brickley, M., & McKinley, J. I. (2017). In M. Brickley & J. I. McKinley (Eds.), *Update Guidelines to the Standards for Recording Human Remains* (2017th ed.). INSTITUTE OF FIELD ARCHAEOLOGISTS PAPER NO. 7. British Association for Biological Anthropology and Osteoarchaeology (BABA0). <https://www.babao.org.uk/>
- Buikstra, J. E., & Ubelaker, D. H. (1994). *Standards for data collection from human skeletal remains*. Arkansas Archeological Survey research series, no. 44. (p. 1994). Arkansas Archeological Survey.
- Coles, B., Tyrer, F., Hussein, H., Dhalwani, N., & Khunti, K. (2021). Development, content validation, and reliability of the Assessment of Real-World Observational Studies (ArRoWS) critical appraisal tool. *Annals of Epidemiology*, 55, 57–63.e15. <https://doi.org/10.1016/j.annepidem.2020.09.014>
- Denman, D., Kim, J. H., Munro, N., Speyer, R., & Cordier, R. (2019). Describing language assessments for school-aged children: A Delphi study. *International Journal of Speech-Language Pathology*, 21(6), 602–612. <https://doi.org/10.1080/17549507.2018.1552716>
- Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, 67(4), 401–409. <https://doi.org/10.1016/j.jclinepi.2013.12.002>
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (4th ed.) (p. 528). John Wiley & Sons Inc.
- Donoghue, P. C. J., & Rücklin, M. (2016). The ins and outs of the evolutionary origin of teeth. *Evolution & Development*, 18(1), 19–30. <https://doi.org/10.1111/ede.12099>
- Enthoven, W. T. M., Roelofs, P. D. D. M., Deyo, R. A., van Tulder, M. W., Koes, B. W., & Cochrane Back and Neck Group. (2016). Non-steroidal anti-inflammatory drugs for chronic low back pain. *Cochrane Database of Systematic Reviews*, 2(2), CD012087. <https://doi.org/10.1002/14651858.CD012087>
- Evensen, J. P., & Øgaard, B. (2007). Are malocclusions more prevalent and severe now? A comparative study of medieval skulls from Norway. *American Journal of Orthodontics and Dentofacial Orthopedics*, 131(6), 710–716. <https://doi.org/10.1016/j.ajodo.2005.08.037>
- Fladmark, K. R. (1978). *A guide to basic archaeological field procedures*, Publication No. 4 (p. 1978). Dept. of Archaeology, Simon Fraser University.
- Goodwin, P., Ord, J. K., Öller, L. E., Sniezek, J. A., & Leonard, M. (2002). Principles of Forecasting: A Handbook for Researchers and Practitioners. *International Journal of Forecasting*, 18(3), 468–478. [https://doi.org/10.1016/S0169-2070\(02\)00034-1](https://doi.org/10.1016/S0169-2070(02)00034-1)
- Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, 32(4), 1008–1015. <https://doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x>
- Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savovic, J., Schulz, K. F., Weeks, L., Sterne, J. A., Cochrane Bias Methods Group, & Cochrane Statistical Methods Group. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343(oct18 2), d5928. <https://doi.org/10.1136/bmj.d5928>
- Holloway, K., Henneberg, R. J., de Barros Lopes, M., & Henneberg, M. (2011). Evolution of human tuberculosis: A systematic review and meta-analysis of paleopathological evidence. *Homo*, 62(6), 402–458. <https://doi.org/10.1016/j.jchb.2011.10.001>
- Hong, C. S. (2010). Relationship Between Patient Panel Characteristics and Primary Care Physician Clinical Performance Rankings. *JAMA*, 304(10), 1107–1113. <https://doi.org/10.1001/jama.2010.1287>
- Hsu, C.-C., & Sandford, B. A. (2007). The Delphi Technique: Making Sense of Consensus. *Practical Assessment, Research, and Evaluation*, 12(1), 10. <https://doi.org/10.7275/pdz9-th90>
- Jackes, M. (2011). Representativeness and Bias in Archaeological Skeletal Samples. In *Social Bioarchaeology* (pp. 107–146). Wiley-Blackwell. <https://doi.org/10.1002/9781444390537.ch5>
- Kaidonis, J. A. (2008). Tooth wear: the view of the anthropologist. *Clinical Oral Investigations*, 12(S1), 21–26. <https://doi.org/10.1007/s00784-007-0154-8>
- Lavelle, C. L. B. (1972). A comparison between the mandibles of Romano-British and nineteenth century periods. *American Journal of Physical Anthropology*, 36(2), 213–219. <https://doi.org/10.1002/ajpa.1330360209>
- Mahajan, V., Linstone, H. A., & Turoff, M. (1976). The Delphi Method: Techniques and Applications. *Journal of Marketing Research*. Edited by H.A. Linstone and M. Turoff, 13(3), 317–318. <https://doi.org/10.2307/3150755>
- Mays, S., Brickley, M., & Dodwell, N. (2004). *Human Bones from Archaeological Sites: Guidelines for Producing Assessment Documents and Analytical Reports*, English Heritage Publications. Edited by D. M and Jone. English Heritage.
- Mokkin, L. B., Boers, M., van der Vleuten, C. P. M., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2020). COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Medical Research Methodology*, 20(1), 293–306. <https://doi.org/10.1186/s12874-020-01179-5>

- Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., Bouter, L. M., & de Vet Henrica, C. W. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology*, *10*(1), 22. <https://doi.org/10.1186/1471-2288-10-22>
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745. <https://doi.org/10.1016/j.jclinepi.2010.02.006>
- Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2018). Scale development: ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, *30*(1), 3. <https://doi.org/10.1186/s41155-016-0057-1>
- Murphy, M. K., Black, N. A., Lamping, D. L., McKee, C. M., Sanderson, C. F., Askham, J., & Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development. *Health technology assessment (Winchester, England)*, *2*(3), 1–88. <https://doi.org/10.3310/hta2030>
- Nikita, E., & Karligiotti, A. (2019). *BASIC GUIDELINES FOR THE EXCAVATION AND STUDY OF HUMAN SKELETAL REMAINS*. The Cyprus Institute Science and Technology in Archaeology and Culture Research Center (STARC).
- OSSAFreelance. (2005). *A Field Guide to the Excavation of Inhumated Human Remains, British Archaeological Jobs Resources, Practical Guide 14*. Available at: <http://www.bajr.org/BAJRGuides/14>. Field Guide to the Excavation of Human Inhumated Remains/FieldGuidetotheExcavationofHumanInhumatedRemains.pdf.
- Quam, R., Bailey, S., & Wood, B. (2009). Evolution of M 1 crown size and cusp proportions in the genus Homo. *Journal of Anatomy*, *214*(5), 655–670. <https://doi.org/10.1111/j.1469-7580.2009.01064.x>
- Rajbhoj, A. A., Parchake, P., Begnoni, G., Willems, G., & Cadenas de Llano-Pérula, M. (2021). Dental changes in humans with untreated normal occlusion throughout lifetime: A systematic scoping review. *American Journal of Orthodontics and Dentofacial Orthopedics*, *160*(3), 340–362.e3. <https://doi.org/10.1016/j.ajodo.2021.02.014>
- Rangel, S. J., Kelsey, J., Colby, C. E., Anderson, J. D., & Moss, R. L. (2003). Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. *Journal of Pediatric Surgery*, *38*(3), 390–396. <https://doi.org/10.1053/jpsu.2003.50114>
- Scott, G. R. (2018). Dental Anthropology. In *Encyclopedia of Global Archaeology* (pp. 1–8). Springer International Publishing. https://doi.org/10.1007/978-3-319-51726-1_138-2
- Slim, K., Nini, E., Forestier, D., Kwiatkowski, F., Panis, Y., & Chipponi, J. (2003). Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. *ANZ Journal of Surgery*, *73*(9), 712–716. <https://doi.org/10.1046/j.1445-2197.2003.02748.x>
- Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., ... Higgins, J. P. T. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, i4919. <https://doi.org/10.1136/bmj.i4919>
- Streiner, D. (2003). Clinimetrics vs. psychometrics: an unnecessary distinction. *Journal of Clinical Epidemiology*, *56*(12), 1142–1145. <https://doi.org/10.1016/j.jclinepi.2003.08.011>
- Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research*, *27*(5), 1159–1170. <https://doi.org/10.1007/s11136-018-1829-0>
- von der Gracht, H. A. (2012). Consensus measurement in Delphi studies. *Technological Forecasting and Social Change*, *79*(8), 1525–1536. <https://doi.org/10.1016/j.techfore.2012.04.013>
- von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2014). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. *International Journal of Surgery*, *12*(12), 1495–1499. <https://doi.org/10.1016/j.ijsu.2014.07.013>
- Walker, P. L., Johnson, J. R., & Lambert, P. M. (1988). Age and sex biases in the preservation of human skeletal remains. *American Journal of Physical Anthropology*, *76*(2), 183–188. <https://doi.org/10.1002/ajpa.1330760206>
- Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. M. M., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, *3*(1), 25–38. <https://doi.org/10.1186/1471-2288-3-25>
- Whiting, P., Wolff, R., Mallett, S., Simera, I., & Savović, J. (2017). A proposed framework for developing quality assessment tools. *Systematic Reviews*, *6*(1), 204–213. <https://doi.org/10.1186/s13643-017-0604-6>
- Wood, B. (1996). Human evolution. *BioEssays*, *18*(12), 945–954. <https://doi.org/10.1002/bies.950181204>
- Wright, J. (1992). A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *Journal of Clinical Epidemiology*, *45*(11), 1201–1218. [https://doi.org/10.1016/0895-4356\(92\)90161-F](https://doi.org/10.1016/0895-4356(92)90161-F)
- Yoon, S., Speyer, R., Cordier, R., Aunio, P., & Hakkarainen, A. (2020). A Systematic Review Evaluating Psychometric Properties of Parent or Caregiver Report Instruments on Child Maltreatment: Part 2: Internal Consistency,

Reliability, Measurement Error, Structural Validity, Hypothesis Testing, Cross-Cultural Validity, and criterion validity. *Trauma, Violence, & Abuse*, 22, 1296–1315. <https://doi.org/10.1177/1524838020915591>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Rajbhoj, A. A., Speyer, R., Crevecoeur, I., Begnoni, G., Willems, G., & Cadenas de Llano-Pérula, M. (2022). Development of a tool to assess the methodological quality of studies reporting on archaeologically excavated human skeletons: An international Delphi study. *Archaeometry*, 1–15. <https://doi.org/10.1111/arcm.12786>